本文引用格式：陈庆端.一种 CLIP 自监督学习的多模态睡眠分期方法[J].自动化与信息工程,2024,45(4):24-29;35.
CHEN Qingduan. A multi modal sleep staging method of CLIP self supervised learning[J]. Automation & Information Engineering, 2024,45(4):24-29;35.

一种 CLIP 自监督学习的多模态睡眠分期方法*

陈庆端
（广东工业大学，广东 广州 510006）

摘要：睡眠分期对睡眠质量评估、睡眠障碍诊断具有重要的意义。针对基于深度学习的睡眠分期存在标签数据少、数据标注困难等问题，提出一种 CLIP 自监督学习的多模态睡眠分期方法。通过学习无标签数据的特征表示，解决了因标签数据少而导致的模型训练效果欠佳的问题。在不同标签数据下，将基于 CLIP 的多模态自监督学习方法与有监督学习、单模态自监督学习方法 SimCLR 和 TS-TCC 进行对比实验。实验结果表明，基于 CLIP 的多模态自监督学习方法能有效提高睡眠分期的性能。

关键词：多模态自监督学习；睡眠分期；CLIP；单模态自监督学习；有监督学习
中图分类号：TN911.7; R318　　**文献标志码：**A　　**文章编号：**1674-2605(2024)04-0004-07
DOI：10.3969/j.issn.1674-2605.2024.04.004　　开放获取

A Multi Modal Sleep Staging Method of CLIP Self Supervised Learning

CHEN Qingduan
(Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Sleep staging is of great significance for assessing sleep quality and diagnosing sleep disorders. A multi modal sleep staging method of CLIP self supervised learning is proposed to address the problems of limited labeled data and difficulty in data annotation in deep learning based sleep staging. By learning the feature representation of unlabeled data, the problem of poor model training performance caused by limited labeled data has been solved. Comparative experiments will be conducted between CLIP based multi modal self supervised learning method and supervised learning, single modal self supervised learning methods SimCLR and TS-TCC under different labeled data. The experimental results indicate that the multi modal self supervised learning method based on CLIP can effectively improve the performance of sleep staging.

Keywords: multi modal self supervised learning; sleep staging; CLIP; single modal self supervised learning; supervised learning

0　引言

睡眠是人类必不可少的生理活动，睡眠质量与人的健康密切相关[1]。当睡眠不足或睡眠质量不佳时，可能会引起自主神经紊乱，影响人体代谢、免疫和内分泌机能，增加罹患生理系统功能障碍的风险，如失眠、记忆力下降、睡眠呼吸暂停等[2-3]。随着睡眠障碍的愈发普遍，睡眠质量的监测、评估和睡眠相关疾病的诊断变得尤为重要，而睡眠分期是睡眠质量评估的前提和睡眠相关疾病诊断的基础。

近年来，随着人工智能技术的发展，深度学习方法，如基于卷积神经网络（convolutional neural network, CNN）和循环神经网络（recurrent neural network, RNN）等监督学习方法，广泛应用于睡眠分期领域，提高了睡眠质量监测和评估的效率。MOUSAVI 等[4]利用 SleepEEG-Net 深度卷积神经网络来提取时间序列的内部特征，在 20 折交叉验证下，准确率为 84.26%、F1 分数为 79.66%。PHAN 等[5]提出 SeqSleepNet 循环神经网络，加强了睡眠帧间上下文的关联性获取。

* 基金项目：国家自然科学基金项目（62073086）

SUPRATAK 等[6]结合 CNN 和 RNN 提出 DeepSleepNet，不仅能提取睡眠帧的脑电图（electroencephalogram, EEG）波形特征，还能学习相邻睡眠帧间的转换规律。SUN 等[7]和 PRADEEPKUMAR 等[8]将 EEG 和眼电图（electrooculogram, EOG）作为网络输入，利用多模态自监督学习方法提高了睡眠分期的性能。但利用这些监督学习方法进行睡眠分期，需要大量的睡眠生理信号数据，这些数据的收集和标注需耗费大量的时间和人力成本，且难以获得数量充足的高质量数据来训练模型，制约了现有睡眠分期模型性能的进一步提高。

自监督学习利用无标签数据进行预训练，为下游任务学习通用的表征，从而引导模型训练，是解决标签数据少的有效方法。现有的自监督学习多应用于自然语言处理和计算机视觉领域，但睡眠分期常用的生理时间序列信号是一种低维数据，具有非线性、非平稳性，且相邻或一定间距的数据帧之间存在较强的相关性。

本文将基于对比语言-图像预训练（contrastive language-image pretraining, CLIP）[9]的多模态自监督学习应用于睡眠分期，提出一种 CLIP 自监督学习的多模态睡眠分期方法，解决了标签数据少、数据标注困难的问题。

1 相关内容

1.1 自监督学习

自监督学习的本质是构建辅助任务和伪标签来学习无标签数据的特征表示。根据输入信息的形式，自监督学习可分为单模态自监督学习和多模态自监督学习。其中，单模态自监督学习利用单一的模态数据构建辅助任务，从无标签数据中挖掘监督信息。典型的单模态自监督学习方法 SimCLR[10]，将每条样本和对应的数据增强样本视为正对，同一训练批次内的其他样本视为负对，让互为正对的样本相互吸引，互为负对的样本相互排斥。多模态自监督学习不仅为下游任务学习通用表征而构建辅助任务，还学习各个模态信息间的匹配关系[11]，合理处理多模态信息可让学习到的特征表示更加全面准确，进一步提高模型的泛化能力和下游任务的性能。典型的多模态自监督学习方法 CLIP，将文本和图像进行匹配，在预训练时通过简单的预测配对，实现较好的零次性能。LI 等[12]提出融合前对齐（align before fuse, ALBEF）的自监督学习方法，在视觉和文本特征输入 Transformer 前，先对齐，再融合来学习多模态表征；BAEVSKI 等[13]提出的 Data2Vec2.0 将语音、视觉、文本 3 个模态数据通过框架整合，提高了计算效率。

自监督学习根据辅助任务类型可分为生成式自监督学习和对比式自监督学习[14]。相较于生成式自监督学习，对比式自监督学习不需要对原始数据进行重构，只要求模型能够区分相似样本与不相似样本，侧重于学习样本间的共同特征，且模型的泛化能力更强。因此，本文选用对比式自监督学习进行自监督预训练。

1.2 睡眠分期

睡眠分期根据人类睡眠过程中的脑电表现、眼球运动情况和肌肉张力的变化，将睡眠过程分为不同的阶段。根据美国睡眠医学学会的标准[15]，人的睡眠被分为 W、N1、N2、N3 和 REM 5 个阶段。剔除标签为 UNKNOW 和 MOVEMENT 的数据后，将其余标签的数据以 30 s 为一帧进行分段，每条片段对应一个睡眠阶段。ELDELE 等[16]提出基于时间和上下文对比的时间序列表征学习框架（time-series representation learning framework via temporal and contextual contrasting, TS-TCC）的自监督学习方法，在 3 个公开数据集上，仅用 5%~10%的有标签数据就达到与有监督学习使用 100%有标签数据相当的分类效果。YE[17]提出了 CoSleep，通过多视图学习协同训练学习 EEG 的表征信息，在 Sleep-EDF20 数据集上，使用 10%的有标签数据，达到了 71.9%的准确率（accuracy, ACC）。XIAO[18]提出了 SleepDPC，结合预测对比编码和鉴别编码，从原始 EEG 中学习高级语义，使用 10%的有标签数据，达到了 70.1%的 ACC。CHANG[19]提出深序列睡眠网络（deep sequential sleep network, DSSNet），通过学习单通道 EEG 的多视图表示和最大化负样本之间的相关性，取得了比 CoSleep 和 SleepDPC 更好的睡眠分期性能。

2 睡眠分期模型

本文睡眠分期模型的主干网采用基于 U²-Net[20] 的 U 型编解码网络架构。U²-Net 的基本单元是类似 U-Net[21]结构的 U 型单元，能够从不同尺度提取特征并进行特征融合。本文选择的特征提取网络 U²-Net 结构与 SalientSleepNet[22]相似，如图 1 所示。

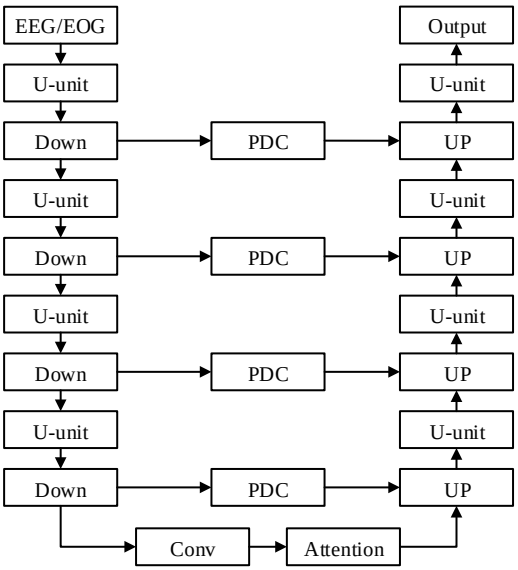

图 1　U²-Net 基本架构[20]

图 1 中，Conv 模块包括卷积层、BN 层和激活函数 ReLU[23]。为了同时学习 EEG/EOG 信号中不同距离的时序特征，在编码器与解码器之间构建了并行空洞卷积（parallel dilated convolution, PDC）模块，包含 4 个不同膨胀率（1、2、3、4）的卷积层。

睡眠生理信号特征提取神经网络的输入是经过预处理的 Fpz-Cz 单通道 EEG/EOG 信号。

首先，EEG/EOG 信号分别经过独立的 U²-Net 提取各自的睡眠波性特征，网络中嵌套的 U 型结构可以在不同层次有效地提取 EEG/EOG 信号的多尺度特征。下采样编码与上采样解码之间的 PDC 模块可以增大网络的感受野，获取多尺度特征，帮助网络学习 EEG/EOG 信号的短距离、长距离时序信息。

然后，采用注意力模块来抑制时间序列中无关的特征信息，突出重要的局部特征信息，学习模态内的上下文关系。

接着，将 EEG/EOG 网络的输出结果进行特征融合：

$$X_{\text{fuse}} = X_1 + X_2 + X_1 \cdot X_2 \tag{1}$$

式中：X_1 和 X_2 分别为 EEG/EOG 信号经过 U²-Net 解码器的输出。

最后，融合后的特征通过 Solfmax 分类器将睡眠波性特征映射到 5 个睡眠阶段标签上，以实现睡眠分期。

3 多模态自监督学习方法 CLIP

CLIP 主要用于图文多模态训练，旨在通过自监督的方式让睡眠分期模型理解图像。CLIP 构建大量的图像-文本对进行训练，使睡眠分期模型在向量空间中将相应的图像-文本对相近，不相应的图像-文本对偏离。

3.1 样本对构建

在基于 CLIP 自监督学习的多模态睡眠分期方法中，EEG/EOG 信号作为两种不同的模态，在同一批次的 N 条样本中，首先，EEG/EOG 信号分别经过图 1 网络的下采样编码过程进行特征编码并归一化，得到表征 I_1，I_2,⋯,I_N 和 T_1，T_2,⋯,T_N；然后，构建正负样本对，将来自同一样本的具有相关性的两种模态信息视为正样本对(I_j, T_j)，将来自不同样本的相关性不强的两种模态信息视为负样本对(I_j, T_k)；最后，计算相应的相似度。为了使学习到的特征具有较高的模态相关性，在嵌入空间中将正样本对的表示距离拉近，负样本对的表示距离拉远。因此，需要最大化正样本对的相似度，最小化负样本对的相似度。

3.2 损失函数

为了学习 EEG/EOG 两种模态的数据表征，且保证相关联模态数据间的距离尽可能近，不相关联模态数据间的距离尽可能远，损失函数表达式定义为

$$L = \frac{1}{2}\left(-\log\frac{\exp(sim(I_i,T_i)/\tau)}{\sum_{k=1}^{2N}\exp(sim(I_i,T_k)/\tau)} - \log\frac{\exp(sim(T_i,I_i)/\tau)}{\sum_{k=1}^{2N}\exp(sim(T_i,I_k)/\tau)}\right), k \neq i \tag{2}$$

式中：*sim* 为 2 个模态数据间的相似度，具体为交叉熵损失；*I* 和 *T* 分别为归一化后的 EEG/EOG 特征表示；τ 为温度系数，用于控制睡眠分期模型对负样本的区别度，若 τ 过大，对相似度大的负样本学习的关注度将减小，若 τ 过小，可能导致模型难以收敛或泛化性能下降，本文设置 $\tau = 0.06$。

损失函数为对称的相似度计算：在同一批次的 EEG 和 EOG 特征表示中，为每条 EEG 特征表示寻找匹配的 EOG 表征；同样地，为每条 EOG 特征表示寻找匹配的 EEG 表征。在公式(2)中，通过分子计算正样本对的相似度，分母计算负样本对的相似度。若样本对越相似，则分子越大、分母越小，损失函数也越小。

4 单模态自监督学习方法

4.1 辅助任务构建

单模态自监督学习方法通过数据增强从原始信号中生成两种不同但相关的视图来构造正负样本。本文对每条 30 s 长的 EEG/EOG 信号做两种数据增强：1） 将 EEG/EOG 信号分成若干段并随机打乱顺序；2） 对 EEG/EOG 信号随机裁剪，并采用双线性插值法将其放大为原来的长度（30 s）。对于同一批次的 EEG/EOG 信号特征表示，由相同原数据的另一种增强方式得到的特征表示是其正样本，其他数据是其负样本。计算同一模态内两种不同增强方式所得特征表示的相似度，最大化同一样本不同增强视图之间的相似度，最小化其与其他样本间的相似度，从而让睡眠分期模型学习到更通用的表征。

4.2 SimCLR 和 TS-TCC 方法

1） SimCLR，随机抽取一个小批量样本，对每条样本进行两种数据增强，计算样本对的余弦相似度，使相关样本对互相吸引，不相关样本对互相排斥。

2） TS-TCC，原始时间序列信号利用弱增强（原始时间序列信号加入随机噪声）、强增强（原始时间序列信号分段并打乱顺序，再加入随机噪声）生成两种增强视图，先通过跨视图预测任务学习时间表征，再通过上下文对比，最大化相同样本不同上下文之间的相似性，最小化不同样本上下文之间的相似性。

5 评价指标

本实验采用 ACC、F1 分数（F1-score, F1）作为睡眠分期模型的评价指标，计算公式分别为

$$A_{\mathrm{ACC}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$P_{\mathrm{PRE}} = \frac{TP}{TP + FP} \tag{4}$$

$$R_{\mathrm{REC}} = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2P_{\mathrm{PRE}} \cdot R_{\mathrm{REC}}}{P_{\mathrm{PRE}} + R_{\mathrm{REC}}} \tag{6}$$

式中：*TP* 为真正例，表示分类器将正例正确分为正例的数量；*FP* 为假正例，表示分类器将负例分为正例的数量；*TN* 为真负例，表示分类器将负例正确分为正例的数量；*FN* 为假负例，表示分类器将正例分为负例的数量；P_{PRE} 为精确率；R_{REC} 为召回率。

6 数据

6.1 数据集

本实验数据来源于 PhysioNet 平台的 Sleep-EDF20 数据集和 Sleep-EDF78 数据集[24]，其中 Sleep-EDF20 数据集（有 20 名受试者）是 Sleep-EDF78 数据集（有 78 名受试者）的一个子集。数据采样率为 100 Hz。采集每名受试者两晚的多导睡眠图（polysomnography, PSG）数据（由于数据丢失，少数受试者只有一晚的数据）。PSG 数据包含 EEG 信号、EOG 信号、肌电图（electromyogram, EMG）信号。将 Sleep-EDF78 数据集中不包含 Sleep-EDF20 数据集的部分用于自监督预训练，将 Sleep-EDF20 数据集用于有监督训练。睡眠各阶段 Sleep-EDF20 数据集的样本数量如表 1 所示。

表 1 睡眠各阶段 Sleep-EDF20 数据集的样本数量

睡眠阶段	样本数量/条
W	8 285(19.6%)
N1	2 804(6.6%)
N2	17 799(50.4%)
N3	5 703(12.9%)
REM	7 717(17.9%)
总数	42 308

6.2 数据预处理

为减小数据分布的差异给睡眠分期模型训练带来的影响，提高模型的泛化能力，对每段原始EEG/EOG信号（每段30 s，记为1帧）进行Z-Score标准化处理，使不同数据具有相似的分布。Z-Score标准化公式为

$$\bar{x}=\frac{x-\mu}{\sigma} \tag{7}$$

式中：x为原始EEG/EOG数据帧，μ为EEG/EOG数据帧的均值，σ为EEG/EOG数据帧的方差。

7 对比实验

实验环境：利用Python3.7安装的深度学习框架Keras搭建神经网络，计算机处理器为3.4 GHz Intel Core i7-6800K，内存为NVIDIA GeForce RTX 2070 24 GB。

本实验选取的优化器为Adam[25]，β_1=0.9，β_2=0.999，学习率为1×10^{-3}，最小批次为4。验证集和测试集数据各来自Sleep-EDF20数据集中4名不同的受试者，其余的12名受试者数据为训练集。为验证本文方法在不同的受试者数量有标签数据下的训练效果，按受试者数量设置4组对比实验，每组分别包含3名受试者、6名受试者、9名受试者和12名受试者，且受试者数量多的组包含了受试者数量少的组的数据。

将自监督学习的预训练权重迁移到下游的分类任务中进行微调，对比有监督学习、SimCLR、TS-TCC、CLIP 4种方法预训练后的性能指标，结果分别如表2、3所示，绘制其相应的曲线，如图2、3所示。其中N为有标签数据来自的个体数量。

表2 4种监督学习方法分类ACC比较 %

方法	$N=3$	$N=6$	$N=9$	$N=12$
有监督学习	61.18	75.36	78.88	82.11
SimCLR	73.50	80.76	81.00	82.71
TS-TCC	73.35	79.81	82.93	83.70
CLIP	73.62	80.82	83.02	84.44

表3 4种监督学习方法分类F1比较 %

方法	$N=3$	$N=6$	$N=9$	$N=12$
有监督学习	56.12	68.69	72.38	75.88
SimCLR	68.33	73.62	73.70	76.87
TS-TCC	65.80	71.90	74.70	76.85
CLIP	68.71	74.14	76.62	78.26

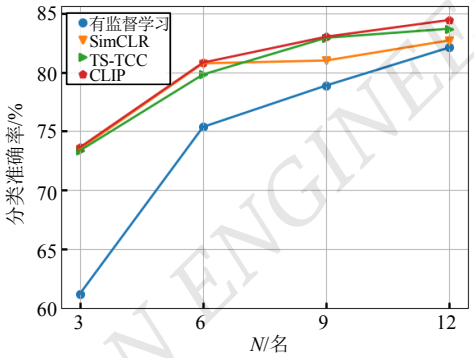

图2 4种监督方法分类ACC(%)曲线

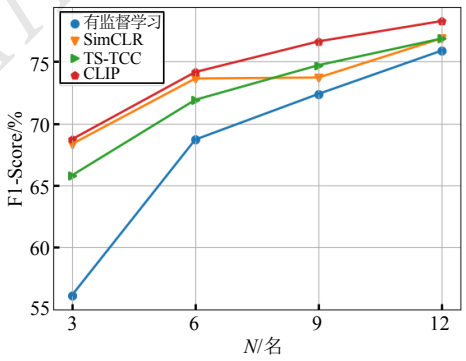

图3 4种监督方法分类F1(%)曲线

由图2、3可以看出：自监督学习方法的睡眠分期性能均优于有监督学习方法，且在有标签数据数量较少时，自监督学习方法对睡眠分期性能的提升更明显；相较于SimCLR、TS-TCC在单模态基础上通过数据增强构建正负样本对，本文方法（CLIP）让一种模态与另一模态构建正负样本对，使预训练模型学习了多模态特征信息，在4组不同受试者数量的实验中均取得了更好的分期性能，说明多模态间的互补性学习更有利于促进下游任务训练的表征。

8 结论

本文基于U²-Net的网络模型，在睡眠公开数据

集 Sleep-EDF20 上划分 4 组不同数量的有标签数据上进行实验，利用基于 CLIP 自监督学习的多模态睡眠分期方法，与有监督学习、单模态自监督学习方法 SimCLR、TS-TCC 进行对比实验。实验结果表明：经过自监督预训练的分期结果优于未经预训练的分期结果；在少标签数据的情况下，自监督学习分类性能的提升更明显，说明自监督预训练能从无标签数据中有效学习到特征表示，促进了下游任务的训练，验证了本文方法在睡眠分期领域的有效性。

参考文献

[1] KOHN TP, KOHN JR, HANEY NM, et al. The effect of sleep on men's health[J]. Transl Androl Urol, 2020,9(S2):S178-S185.

[2] FINAN P H, QUARTANA P J,REMENIUK B, et al. Partial sleep deprivation attenuates the positive affective system: Effects across multiple measurement modalities[J]. Sleep, 2017, 40(1):1-9.

[3] WULFF K, GATTI S, WETTSTEIN J G, et al. Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease[J]. Nature Reviews Neuroscience, 2010,11(8):589-599.

[4] MOUSAVI S, AFGHAH F, Acharya U R. SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach[J]. Plos One, 2019,14(5):e216456.

[5] PHAN H, ANDREOTTI F, COORAY N, et al. SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2019,27(3): 400-410.

[6] SUPRATAK A, DONG H, WU C, et al. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2017,25(11):1998-2008.

[7] SUN C, CHEN C, FAN J, et al. A hierarchical sequential neural network with feature fusion for sleep staging based on EOG and RR signals[J]. Journal of Neural Engineering, 2019,16(6): 066020.

[8] PRADEEPKUMAR J, ANANDAKUMAR M, KUGATHASAN V, et al. Towards interpretable sleep stage classification using cross-modal transformers[J]. arXiv preprint arXiv:2208.06991, 2022.

[9] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]. International Conference on Machine Learning. PMLR, 2021:8748-8763.

[10] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations [C]. International Conference on Machine Learning. PMLR, 2020:1597-1607.

[11] ZONG Y, MAC AODHA O, HOSPEDALES T. Self-supervised multimodal learning: A survey[J]. arXiv preprint arXiv:2304.01008, 2023.

[12] LI J, SELVARAJU R, GOTMARE A, et al. Align before fuse: Vision and language representation learning with momentum distillation[J]. Advances in Neural Information Processing Systems, 2021, 34: 9694-9705.

[13] BAEVSKI A, BABU A, HSU W N, et al. Efficient self-supervised learning with contextualized target representations for vision, speech and language[C]. International Conference on Machine Learning. PMLR, 2023: 1416-1429.

[14] LIU X, ZHANG F, HOU Z, et al. Self-supervised learning: generative or contrastive[J]. IEEE Transactions on Knowledge and Data Engineering, 2021,35(1):857-876.

[15] BERRY R B, BROOKS R, GAMALDO C E, et al. The AASM manual for the scoring of sleep and associated events[Z]. Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine, 2012,176 (2012):7.

[16] ELDELE E, RAGAB M, CHEN Z, et al. Self-supervised contrastive representation learning for semi-supervised time-series classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023,45(12):15604-15618.

[17] YE J, XIAO Q, WANG J, et al. Cosleep: A multi-view representation learning framework for self-supervised learning of sleep stage classification[J]. IEEE Signal Processing Letters, 2021, 29: 189-193.

[18] XIAO Q, WANG J, YE J, et al. Self-supervised learning for sleep stage classification with predictive and discriminative contrastive coding[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 1290-1294.

（下转第 35 页）